DOCUMENT RESUME

ED 280 413 HE 020 264

AUTHOR Ehrich, Elizabeth A.; Michels, Gabriele A.
TITLE Stagewise Regression--A Different Approach to Salary Equity Studies. AIR 1986 Annual Forum Paper.
PUB DATE Jun 86
NOTE 24p.; Paper presented at the Annual Forum of the Association for Institutional Research (26th, Orlando, FL, June 22-25, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Evaluation Methods; Females; Higher Education; Males; Predictor Variables; *Regression (Statistics); *Research Methodology; *Salary Wage Differentials; *Sex Differences
IDENTIFIERS *AIR Forum

ABSTRACT

The use of stagewise regression was investigated in a university study of salary equity for full-time staff-exempt personnel. The commonly used method of regression substitution was modified in three ways: (1) shifting the emphasis away from estimating inequity and toward describing the impact of men's and women's population differences on salaries; (2) using stagewise regression instead of multiple regression; and (3) adding a graphical procedure (plots) to the variable selection process. The basic idea in stagewise regression is that at each step the effects of the most important remaining variables are estimated and removed. Study variables included: age when hired, years of service, years in rank, educational level attained, staff grade, and method of job entry (e.g., new hire, promotion). The stagewise regression method allowed for nearly parallel models for men and women to be developed. It also provided a different perspective from least squares by attributing more of the variation in the response variable to those variables that are judged to be more important. In addition, the agreement, or lack of it, between stagewise and least squares results can be an indicator of how balanced or unbalanced the data are. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED280413

78H

Stagewise Regression--A Different Approach to Salary Equity Studies

By

Elizabeth A. Ehrich
Statistical Analyst
The Pennsylvania State University
304 Old Main
University Park, PA 16802

and

Gabriele A. Michels
Research and Planning Associate
The Pennsylvania State University
304 Old Main
University Park, PA 16802

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Contributed Paper Presented at the
26th Annual Forum
The Association for Institutional Research
Orlando, Florida
June 22-25, 1986

HE 020 264

AIR
for Management Research, Policy Analysis, and Planning

This paper was presented at the Twenty-Sixth Annual Forum of the Association for Institutional Research held at the Marriott Orlando World Center, Orlando, Florida, June 22-25, 1986. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Ann K. Dickey, Chair
Forum Publications Editorial
Advisory Committee

Abstract

Regression methods represent powerful tools for studying salary equity, but rest on assumptions that are not always met by the available data. In an exploratory study of staff salaries at one institution, the commonly used method of regression substitution was modified in three ways: (a) shifting the emphasis away from estimating inequity and toward describing the impact of men's and women's population differences on salaries; (b) using stagewise regression instead of multiple regression; and (c) adding a graphical procedure to the variable selection process.

Stagewise Regression--A Different Approach to Salary Equity Studies

Introduction

Regression methods, including regression substitution, are among the most powerful statistical techniques available for studying salary equity (Sobel, 1983). When the data set includes a complete set of variables which legitimately affect salaries, regression substitution can produce an accurate measure of inequity; the effects of differences in the men's and women's populations on the overall salary discrepancy can be obtained as a by-product. However, the validity of the conclusions rests on assumptions which are not always met by the available data (Allard, 1984, McCabe, 1980, Ferber and Green, 1982, Madden, 1982, Waters and Milliken, 1985). Consequently, unless an ideal data set is used, no strong statement can be made about the degree to which inequity is a factor in the salary structures for men and women. However, even if the data set is not complete, as is usually the case, it is still possible to use the regression approach.

The regression substitution method can be broken into three parts: (a) variable selection; (b) formulating the operative question "what would the women earn if they were paid at the same rate as men?"; and (c) the calculations, using multiple regression, to answer the operative question. In our exploratory study of salaries of staff-exempt personnel, we found it useful to modify

all three of these parts. Since the data set was inadequate for making any strong statement about equity, a more direct way for determining the contributions of different characteristics of the men's and women's populations to the overall difference in their average salaries was pursued.

Classical regression models present some potentially serious shortcomings for descriptive studies. The stagewise regression procedure used here provides an alternative that is both efficient and simple to use. It can also be used to select variables for least squares regression models, and to help assess the results.

The Data Base

The source of data was the University's payroll file. The population used in the analysis included all staff-exempt personnel with full-time appointments of more than six months.

Variables. The following variables were considered for use in the analysis because of their possible effect on salaries:

-- Age
-- Campus (main campus or one of the other 21 locations)
-- Staff Grade (grade 4 is lowest, grade 10 highest)
-- Years of service to the University
-- Years in current rank
-- Education level (coded with one of 11 possible values)
-- Type of contract (also referred to as salary plan. Three types of contracts are coded for staff-exempt personnel: twelve-month standing appointment, standing appointment of less than twelve months, and fixed term

appointments of varying lengths; for the purpose of this analysis, all salaries have been converted to their 12-month equivalents)

-- Method of entry (i.e., mode by which employee entered the present position; four possible methods have been coded: hired from outside the University, promoted within the same unit, transferred from another unit, and reappointed)

-- Equal Employment Opportunity (EEO) code

-- Organizational unit (identified by one of 57 personnel representative codes)

-- Market value (identifying three special categories of employees: engineers, physicians, and data processing personnel).

Data base limitations. The data base has some inherent limitations which could not be overcome in any simple way: the University uses a seven grade classification system for all staff members, but there is no uniform classification system for the diversity of jobs included in this group. As a result, the effects of market demands cannot be analyzed except in a limited number of cases where a special effort was made to identify data processing personnel, physicians, and engineers. There is also no way to relate salaries to qualifications, experience, or performance. The analysis was further constrained by the fact that staff grade levels do not correspond to an exclusive set of salary ranges.

Methodology

Basic issues. The goal of the study was to determine how much of the difference between men's and women's average salaries could be accounted for by differences in relevant characteristics of the men's and women's populations. The data for staff-exempt personnel raised several issues associated with using statistics as an accounting tool.

The diversity of jobs and the lack of information on individual qualifications, along with a generally unbalanced situation where the members of one category may number in the single digits while there are hundreds in another category--for example physicians vs. data processors in the 'market value' variable--make this data the kind that statisticians such as Draper & Smith (1981, p. 259) warn about:

> When we do regression calculations on unplanned data (that is, data arising from continuing operations and not from a designed experiment) some potentially dangerous possibilities can arise. . . The error in the model may well not be random but may result from the joint effect of several variables not incorporated in the regression equation, nor, perhaps, even measured. . . . Due to the possibilities of bias in the estimates . . . an observed false effect of a visible variable may, in fact, be caused by an unmeasured latent variable. Provided the system

> continues to run in the same way as when the data were recorded, this will not mislead. However, because the latent variable is not measured, its changes will not be seen or recorded, and such changes may well cause the predicted equation to become unreliable. . . . Where designed experiments are not feasible, happenstance data may still be analyzed via regression methods. However, the additional possibilities of jumping to erroneous conclusions must be kept in mind.

In our study we were interested in making generalizations from one population to another, rather than from one time to another, but the principle is the same. Our data also differs from the situation described above in that we actually have data on both groups, which is not the case when trying to predict a future situation. (A more extensive discussion of observational studies can be found in Statistical Methods for Comparative Studies by Anderson, Auquier, Hauck, Oakes, Vandaele and Weisberg, 1980.)

We decided that a variable could be used to 'account for' part of the overall salary differential only if its effect on men's salaries was close to the effect it had on women's salaries. For example, if the men's salaries were $190 higher for each year of service, and the women's were $200 higher per year of service, then a population difference of 3 years of service would 'account for' about $600. On the other hand, if the women's salaries were

only $50 higher for each year of service, and the men's still $190, more questions would be raised than answered. In that case, using the variable 'years of service' in the analysis would not further our goal of accounting for differences, and we would leave it out.

Even when variables are used only when their effects on men's and women's salaries are similar, it must still be decided whether to use the men's regression coefficients or the women's regression coefficients to calculate the 'accounted for' salary differentials. A third option is to use a single regression for men and women combined. For a strict equity study, it is appropriate to use the men's regression equation to answer the question: 'What would the women earn if they were compensated at the same rate as the men, but based on the women's characteristics?'. The salary differentials associated with each variable can be obtained as a by-product of this regression, but if these differentials are the primary interest of the study, this is not necessarily the best way of estimating them.

For a descriptive study where there is no hope of getting an accurate estimate of inequity, it may be better to estimate the 'accounted for' salary differentials directly, as answers to questions like 'how much higher (lower) would the women's average salary be if they had the same average years of service as the men?' We felt that the best answers to questions about women's salaries may be those that are based on women's data, so we used the women's regression equation to calculate the answers.

One key to obtaining reliable answers to the operative questions pertaining to women's salaries as they relate to each of the variables was to estimate the coefficients for each of the variables one at a time, using stagewise regression, instead of simultaneously with least squares regression. While least squares is optimal for predicting new observations, the individual coefficients of the variables may be different from what is expected (Draper and Smith, 1981, p. 340). For example, the least squares method may produce the unlikely result of the physicians earning less than the counselors, while also stating that persons with medical degrees earn more than those with master's degrees. Stagewise regression would yield more realistic results.

Stagewise regression. Looking at variables one at a time implies a series of simple regressions; however, the fact that some variables are correlated means that some of the overall differences in average salaries will be accounted for more than once. Stagewise regression is a hierarchical procedure that estimates marginal, or one at a time, effects like simple regressions, except that the estimates are corrected for correlations with variables used in earlier stages. Consequently, they can be used in a single additive regression equation, like least squares estimates. For perfectly designed (orthogonal) experiments, stagewise regression and multiple regression are identical.

The basic idea in stagewise regression is that at each step, the effects of the most important remaining variable are estimated

and removed. For the first step, the response variable in our example is salaries. The regression model is fitted and the residuals (= salary - predicted salary) are calculated. For the second stage, the residuals are the response variable. The residuals from the second stage (residual - predicted residual) are the response variable for the third, and so on. The equations from each stage are added, to obtain the final stagewise regression equation. (Draper and Smith, 1981, p. 337).

By including the most important variables in the earlier stages, the possibility of less important variables seeming more important than they are, because of confounding or correlation with other variables, is eliminated. Another feature of stagewise regression is that the data analyst can control the order in which variables enter the equation, so that information about the actual structure of the problem can be used to advantage. For example, temperature and number of people at the beach may both be highly correlated with wild mushroom growth. With stagewise regression we could enter temperature, the more logical choice, first. If the number of people at the beach still appears to have an effect on the residuals, we can use this variable as well; however, the coefficient for 'number of people' will be smaller than the one obtained with multiple regression, while the coefficient for 'temperature' will be larger than it would have been with multiple regression.

Another advantage of stagewise regression in our study was that it was easier to make sure that only variables with parallel

effects were included. Once the coefficients for the variable are estimated, they do not change. This is not the case with multiple regression: a variable may seem to have the same effect on men's and women's salaries, but when other variables are added to the model, that pattern could change, and the effects of the variable may no longer appear to be similar. This could present a problem if the original variable is one that we know is important.

In some situations the researcher may want to include variables with non-parallel effects. By using stagewise regression, the 'parallel' variables could be fitted into the model first, and the 'non-parallel' ones later, thus assigning the latter less importance in the model. A similar strategy could be used when we want to include variables which are 'tainted' in some other way.

Variable selection. Although some of the variables in the data set may not obviously be related to salaries, for example type of contract, most of the variables were considered for use in the model anyway. The basic criteria for inclusion were that the variable have some noticeable relationship to salaries, and that the relationship was the same for men as for women. The latter condition was examined with the help of some plots for categorical variables.

The plots represent average salaries, or average residuals for stages past the first, of the men and women in each category of the variable. In the first plot, the variable categories are labeled at the bottom (see Figure 1). It may be difficult to decide if the averages for the men and women are

Figure 1

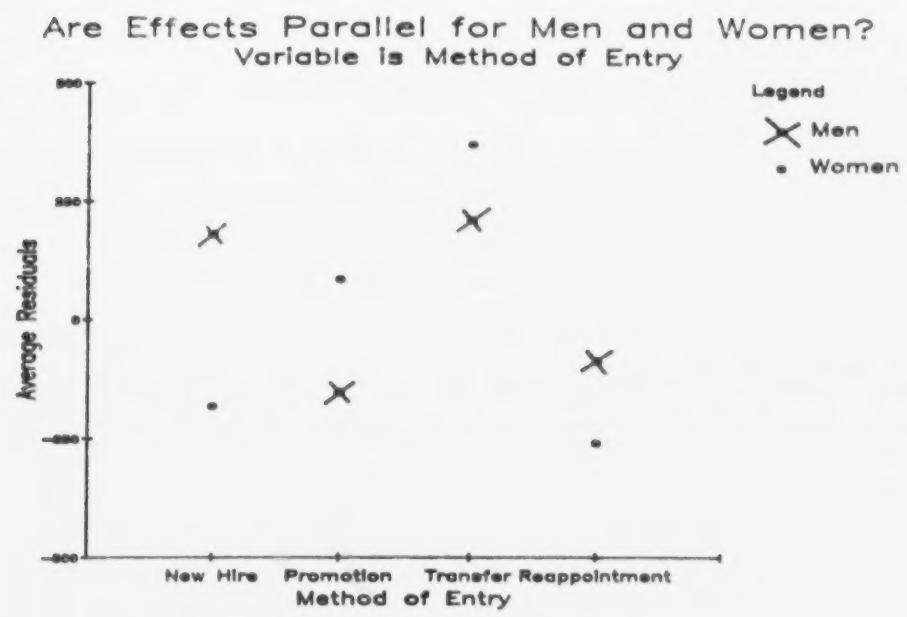
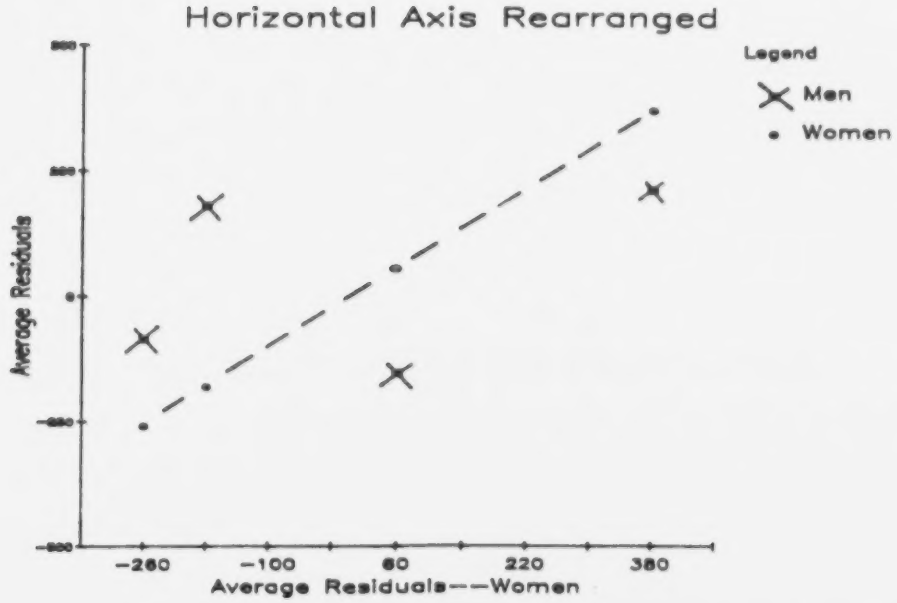


Figure 2



Plotting men's and women's average residuals against categories of a variable does not always reveal whether they follow the same trend (above). When plotted against the women's averages, the women's averages follow a diagonal line--if the effects are parallel, the men's averages will parallel that line.

parallel in this figure, so in Figure 2 both the men's and women's averages for each category are plotted against the women's averages. Now the women's averages fall on the diagonal line, and it is easier to see if the men's averages are parallel to them.

A problem with the plots may occur if some categories have very few observations. It is a good idea to mark the frequencies on the plot next to the points in question, or to blank them out. Some people may prefer to fit a model to data with both sexes, and then test an interaction term for significance. However, it must be remembered that a non-significant interaction is not the same as parallel effects (no interaction). A decision rule should choose the better of the two alternatives (a) closer to parallel or (b) closer to unrelated; ultimately, subjective choices will have to be made.

Plots are especially useful for examining variables that have a large number of categories; in our data, 'organizational unit' had 57 categories. The plots for this variable showed that the men's average salaries in the different organizational units were completely unrelated to the women's salaries in the same units, so the variable was not used in the final model.

Results

Because of the size of the total staff-exempt population, a subset was extracted to construct the examples discussed here. Some sample results, which are consistent with those obtained in the full study, are presented in Tables 1-3. The 12-month equivalent salaries have been disguised as (salary + a)/b, where a

and b are constants. The disguised salaries are referred to as 'compensation', and are measured in 'compensation units'. Table 1 shows the characteristics of the men's and women's populations. Regression models were developed separately for men and women. For comparison with the men's and women's stagewise regression results, two different least squares models were estimated, one using the same variables that were included in the stagewise models, and the other using variables that resulted in 'improved' models with higher r-squared and lower mean squared error. There were thus six models in all, two for each of the regression models.

The coefficients for the six models are shown in Table 2. The stagewise regression estimates (listed in the order in which they were entered in the model) are not startlingly different from the least squares estimates, and in the latter four models the estimates all look reasonable--grade five employees earn more than grade four employees, etc. Because of the order of entering the variables, the coefficients for 'market value' and 'education' are larger in the stagewise model than in the least squares model, while for 'grade' they are smaller.

In the 'improved' models, though, some of the estimates are a little difficult to interpret. In the women's model, the main effect for grade eight is 1,956 compensation units higher than for grade nine. Note, however, that this effect is partially canceled by the especially low value of the coefficient for the interaction variable 'years of service by grade eight'. Still, these estimates suggest that the model is overfitted, that is, more

Table 1

Characteristics of Men's and Women's Populations

Variable	Women's Average or Percent	Men's Average or Percent
Number of Observations	259	414
Compensation Units	$25,232.41	$26,625.76
Age when hired	31.20	30.15
Years of Service	7.95	11.80
Years in rank	3.08	5.25
Education:	100.0%	100.0%
High school	23.9%	18.6%
License or Cert.	0.4%	0.2%
Associate degree	2.7%	12.6%
Bachelor's	47.5%	44.9%
Master's	20.1%	20.3%
Ph.D.	5.0%	2.7%
D.Ed.	0.4%	0.7%
Market Value:	100.0%	100.0%
Data processor	8.5%	12.8%
Not classified	91.5%	87.2%
Staff Grade:	100.0%	100.0%
Four	1.2%	1.0%
Five	25.9%	13.3%
Six	34.7%	27.3%
Seven	30.9%	38.4%
Eight	6.6%	17.6%
Nine	0.8%	2.4%
Method of entry	100.0%	100.0%
New hire	32.4%	28.3%
Promotion	52.5%	50.2%
Transfer	8.5%	15.0%
Re-appointment	6.6%	6.5%

Table 2

Coefficients from Regression Models

Regression Model:	Stagewise		Least Squares		'Improved'	
	Women	Men	Women	Men	Women	Men
R-squared			73%	63%	75%	70%
Root mean squared error			784	1132	755	1047
Variable:						
Intercept	27965	28357	28275	30326	25513	26834
Years of Service	84	96	115	80	444	205
Education:						
High school	-1261	-2974	-478	-1631		
License or Cert.	-1532	-3268	-796	-2458		
Associate degree	-1522	-2565	-629	-1661		
Bachelor's	-1122	-1863	-549	-1507		
Master's	-724	-1545	-331	-1414		
Ph.D.	-998	-708	-550	-1465		
D.Ed.	0	0	0	0		
Market Value:						
Data processor	678	933	230	443	270	657
Not classified	0	0	0	0	0	0
Staff Grade:						
Four	-4130	-4988	-5433	-6021	-2506	-6331
Five	-3346	-3886	-4613	-4764	-1979	-5117
Six	-2608	-3274	-3807	-4018	-1665	-4234
Seven	-1921	-2492	-2854	-3008	-781	-3236
Eight	-223	-1043	-360	-1319	1956	-1484
Nine	0	0	0	0	0	0
Years in rank					33	32
Years of service by grade						
Four					-413	
Five					-381	
Six					-329	
Seven					-313	
Eight					-352	
Nine					0	
Method of entry						
New hire						-292
Promotion						-432
Transfer						41
Re-appointment						0
Age						71
Years service squared						-2
Age by years service						-3

coefficients were estimated than the data can reasonably support. Specifically, in the years of service by grade interaction the relationship between salaries and years of service has been estimated using only two individuals in grade nine, and three in grade four (see Table 1).

Table 3 displays the components that are 'accounted for' by men's and women's population differences, according to the six models. Since the stagewise models were constructed to be nearly parallel, it is not surprising that the components are quite close for the men's and women's equations. Again, the order of fitting in the stagewise procedure results in larger (in absolute value) components being assigned to 'education' and 'market value' than in the least squares.

At the bottom of Table 3 the components have been added to give 'total predicted difference'. Comparing the least squares and stagewise results for the women's models shows a difference of almost 300 compensation units. If the women's data were balanced, as it would be in a designed experiment, the two estimation methods would produce the same results and these quantities would be the same. (Note, however, that the converse is not true: if the quantities are the same, the design is not necessarily balanced.)

In the 'improved' women's model, the total predicted difference is almost as high as the actual difference. One contributor is the relatively high component for (the subtotal) 'years of service', 'years in rank' and 'age when hired'--695 compensation units. Because 'years of service' is overfitted, with

Table 3

Components of Predicted Differences in Men's and Women's Average Compensation

Regression Model	Stagewise		Least Squares		'Improved'	
Variable	Women's Model	Men's Model	Women's Model	Men's Model	Women's Model	Men's Model
Age when hired						-75
Years of Service	321	371	441	307	1707	789
Years service squared						-194
Age by years service						-280
Years of service by grade					-1084	
Years in rank					73	70
Subtotal years of service, years in rank, age	321	371	441	307	695	310
Education	-30	-28	-9	-3		
Market Value	29	40	10	19	12	28
Staff Grade	454	440	620	538	536	564
Method of entry						25
Total Predicted Diff	775	823	1062	862	1243	927
Actual Difference	1393	1393	1393	1393	1393	1393

a separate coefficient for each grade, plus one overall, the sizes of the coefficients are probably bigger than they should be (in absolute value). This is a quirk of least squares regression known as the 'Stein Effect' (Efron and Morris, 1975). The larger implication of the Stein effect for equity studies is that least squares models with too many coefficients for the data can seriously overestimate the total predicted difference (sometimes called the 'total explained' difference).

Discussion

There may be occasions, such as our staff-exempt salary study, when the condition of the data does not warrant spending a lot of time searching for the best regression model. Further, the presence of inescapable confounding may render the results of even technically good models invalid. In our staff-exempt salary study, the stagewise procedure enabled us to develop nearly parallel models for the men and women without undue attention to model fitting intricacies. Stagewise regression also provides a different perspective from least squares by attributing more of the variation in the response variable to those variables which are judged to be more important. In addition, the agreement, or lack of it, between stagewise and least squares results can be an indicator of how balanced or unbalanced the data is.

When least squares estimates are desired, as for court-bound equity studies for example, it is still necessary to avoid using variables that are only significant by virtue of being confounded with more important variables. Stagewise regression used as a

variable selection tool can protect against the use of such variables. On the other hand, it may not guard against overfitting which is caused by including variables with too many categories, such as 'organizational unit' or 'academic department'. The best way to examine variables of this type is with graphical methods; the one discussed above is an example. Another graphical method that can be employed is normal plots which are useful for determining whether category effects are random (Draper and Smith, 1981, p. 177); if they are, they should not be used even if they are highly significant in the regression model.

The hazards caused by variables with too many categories were illustrated for us in the following experiment: the men were randomly divided into two groups, called 'men' and 'women'. A regression equation with terms for 'sex', 'organizational unit' and the interaction of these two variables was calculated three separate times. In one calculation, the variable 'sex' had a significant effect; in another calculation, the interaction term had a highly significant effect. Although the two randomly chosen groups may have really been different in some way, it is more likely that the large number of categories was inducing the 'Stein Effect', resulting in estimates that were larger in absolute value than they should have been.

REFERENCES

Allard, C.A. (1984). Assessing Faculty Salary Equity. The Air Professional File (no. 20). Tallahassee, FL: The Association for Institutional Research.

Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W. and Weisberg, H. I. (1980). Statistical Methods for Comparative Studies. New York: Wiley.

Draper, N. R. & Smith, H. (1981). Applied Regression Analysis. New York: Wiley.

Efron, B., and Morris, C. (1975). Data Analysis Using Stein's Estimator and Its Generalizations. Journal of the American Statistical Association, 70 (350), 311-319.

Ferber, M. A. & Green, C. A. (1982). Employment Discrimination: What Kind of Fairness is Fair? In Preceedings of the American Statistical Association, Social Statistics Section (pp. 159-161.

Madden, J. F. (1982). The Measurement of Employment Discrimination Reduced Forms, Reverse Regression, Comparable Worth and the Definition of Labor Markets. In Proceedings of the American Statistical Association, Social Statistics Section (pp. 162-168.

McCabe, G. P., Jr. (1980). The Interpretation of Regression Analysis Results in Sex and Race Discrimination Problems. The American Statistican, 34(4), 212-215.

Sobel, M. E. (1983). Some Large-Sample Standard Errors for Components of a Mean Difference Under a Linear Model. Sociological Methodology 1983-1984. Leinhardt, S., ed. San Francisco, CA: Jossey-Bass.

Waters, J. & Milliken, G. A. (1985). Evaluating the Effects of Gender in Employment Discrimination Cases: Jurimetric Data Bases, Data Reliability and Strategies for Using Regression Models. In Proceedings of the American Statistical Association, Business Statistics Section (pp. 1-9).